Ninth International Workshop on Image Analysis for Multimedia Interactive Services

3D inference and modelling for video retrieval

Huiyu Zhou, Abdul Sadka and Richard M. Jiang
Brunel University, Uxbridge, Middlesex, United Kingdom
E-mail:{Huiyu.Zhou, Abdul.Sadka, Min.Jiang@brunel.ac.uk}

Abstract

A new scheme is proposed for extracting planar surfaces from 2D image sequences. We firstly perform feature correspondence over two neighboring frames, followed by the estimation of disparity and depth maps, provided a calibrated camera. We then apply iterative Random Sample Consensus (RANSAC) plane fitting to the generated 3D points to find a dominant plane in a maximum likelihood estimation style. Object points on or off this dominant plane are determined by measuring their Euclidean distance to the plane. Experimental work shows that the proposed scheme leads to better plane fitting results than the classical RANSAC method.

1. Introduction

Digital video cameras are widely used, and the quantity of digital videos has dramatically increased recently. For reuse and storage purpose, consumers have to retrieve a video from a large number of multimedia resources. To seek similar videos from a definite database, information retrieval systems have been established with promising performance in searching accuracy and efficiency, e.g. [8]. Many of these established systems attempt to search for videos that have been annotated with metadata *a priori* (e.g. [3]). Nevertheless, there are still a significant number of footages that have been shot but not ever used [1]. These footages normally have not been properly annotated, and hence the retrieval can only be carried out according to the video contents rather than the annotated information.

Of the generic video contents, the need for the ability to retrieve 3D models from databases or the Internet has gained prominence. Content-based 3D model retrieval currently remains a hot research area, and has found its tremendous applications in computer animation, medical imaging, and security. To extract a 3D object, shape-based 3D modelling (e.g. [9]) and similarity or dissimilarity (distance) computation (e.g. [7]) are two of the main research schemes. In this paper, rather than extracting a complete 3D model, we intend to reconstruct flat surfaces from video sequences. This work is inspired by the fact that flat surfaces are one of the basic components of a 3D model, where the estimation of flat surfaces significantly affects the 3D modeling. We believe that the proposed approach in this paper can be used to effectively facilitate the application of 3D model retrieval from databases or Internet in the future.

One of the commonly used strategy to recover flat surfaces is performed using multiple view reconstruction. For example, Bartoli and Sturm [2] used Plucker coordinates to represent the 3D lines in the scope of maximum likelihood estimation, and then they proposed an orthonormal representation to challenge the bundle adjustment problem. Zhou *et al.* [10] conducted coplanarity checks using cross-ratio invariants and periodic analysis of the triangular regions. In this paper, our main contribution is to introduce an iterative RANSAC plane fitting strategy in a maximum likelihood estimation style. This new technique enables us to obtain the best plane fitting to the generated 3D points automatically rather than using empirical criteria that is determined according to a limited number of image samples.

The proposed planar determination algorithm in this paper starts with corner feature detection using two neighboring frames in a monocular video sequence. Given the epipolar geometry constraint, we then build up dense matching between these two groups of points of interest using the sum squared of differences (SSD) correlation method. Assuming a calibrated camera (used to collect this sequence), we then compute a depth map, based on the estimated disparity map. If there is only one single flat surface in the scene, we can launch a RANSAC algorithm [5] to fit a plane to the available three-dimensional points. This RANSAC operation is iterated in an expectation-maximisation context for seeking global minimal errors, which is the main contribution of our work. Note that the proposed strategy works in the presence of motion parallax. To retrieve planes from uncalibrated scenes, we will explore a fast multiple-view reconstruction strategy, based on the algorithm presented in this paper.

978-0-7695-3130-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WIAMIS.2008.37

IEEE computer society

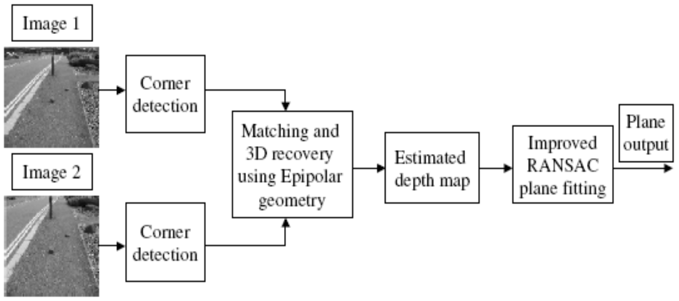


Image 1
Corner detection
Image 2
Corner detection
Matching and 3D recovery using Epipolar geometry
Estimated depth map
Improved RANSAC plane fitting
Plane output

Figure 1. Flowchart of the proposed coplanar determination algorithm.

2 Estimation of a depth map

Before a plane fitting starts, 3D point sets need to be generated based on the 2D image inputs. Of two neighboring images, we consider the later image is the shifted one from the previous image. Given a shift ($\triangle x$,$\triangle y$) and an image point (x,y) in a previous frame, the auto-correlation function for similarity check across frames is defined as $c(x, y) = \sum_W [I(x_i, y_i) - I(x_i + \triangle x, y_i + \triangle y)]^2$, where $I(\cdot)$ denotes the image function and (x_i,y_i) are the image points in the window W (Gaussian) centred at (x,y). The shifted image can be approximated by a Taylor expansion as follows, $I(x_i + \triangle x, y_i + \triangle y) \approx I(x_i, y_i) + [I_x(x_i, y_i), I_y(x_i, y_i)] \begin{bmatrix} \triangle x \\ \triangle y \end{bmatrix}$, where $I_x(\cdot)$ and $I_y(\cdot)$ denote the partial derivations along x and y, respectively. Eventually, we have $c(x, y) = [\triangle x, \triangle y]C(x, y) \begin{bmatrix} \triangle x \\ \triangle y \end{bmatrix}$, where $C(x, y)$ represents the intensity structure of the local neighborhood. Let λ_1 and λ_2 be two eigenvalues of matrix $C(x, y)$. A corner point can be detected if $\min(\lambda_1, \lambda_2)$ is larger than a pre-defined threshold.

Once holding the points of interest, we then apply the sum squared of differences correlation method to match these corner features. Using the matched features, we exploit the well-established epipolar constraints to further refine the correspondence of features. The camera parameters are then used for recovering the scene geometry [10]. As an example, Fig. 2(a) and (b) show the original images superimposed by the extracted corner features using the Harris corner detector [6], (c) is the disparity map and (d) refers to the estimated depth map according to the relationship: $D = fd/z$, where D is depth to be computed, f focal length, d introcular distance and z estimated disparity.

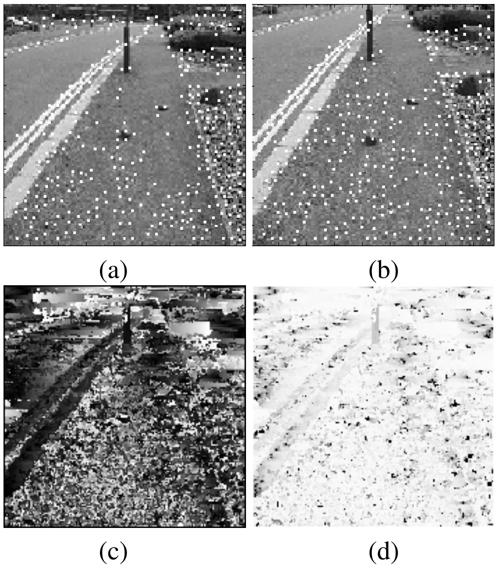
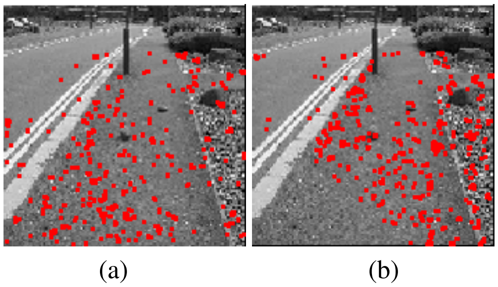


Figure 2. Estimation of disparity and depth maps: (a) and (b) feature extraction, (c) and (d) disparity and depth maps.

3 Iterative RANSAC plane fitting

RANSAC plane fitting is designed to effectively work in the presence of data outliers. This method starts from fitting a plane to a set of 3 points (considered as inliers) randomly selected from the matched corner features. Other image points are then evaluated using the Euclidean distances between these 3D points and the fitted plane. If the points fall in a pre-defined region, then they will be classified as inliers. Otherwise, the points will be removed from the consideration of coplanarity. These steps are repeated until a count limit is reached. In a classical RANSAC plane fitting approach, the iteration is terminated by either a user-specified number or the number of outliers falling below a pre-defined threshold. This heuristic trick cannot handle general situations, where either under- or over-estimation usually appears.

We here intend to find a strategy to achieve maximum likelihood estimation to the flat surfaces. Let N independent samples be represented as $\mathcal{X}$ = $\mathbf{x}_1$, ..., $\mathbf{x}_N$ ($N \geq 30$ denoting a part of the overall image points), the probability density function $p(\mathbf{x})$ (Euclidean distance between the selected 3D points and the fitted plane) and a Gaussian exits as $\mathcal{N}$($\mathbf{x}$,θ,$\mathbf{r}$), where θ and $\mathbf{r}$ stand for a fraction of the inliers of the estimated plane and the relationship between the samples and the inliers, respectively. To obtain a maximum likelihood estimation of θ and $\mathbf{r}$, we can maximise the likelihood function $\Pi_{i=1}^{N} p(\mathbf{x}_i)$. The object function can be



(a) (b)

Figure 3. Estimated ground planes (in red color and hereafter) by (a) the proposed method, and (b) a classical RANSAC technique with the constraint where the number of outliers falls below a pre-defined threshold.

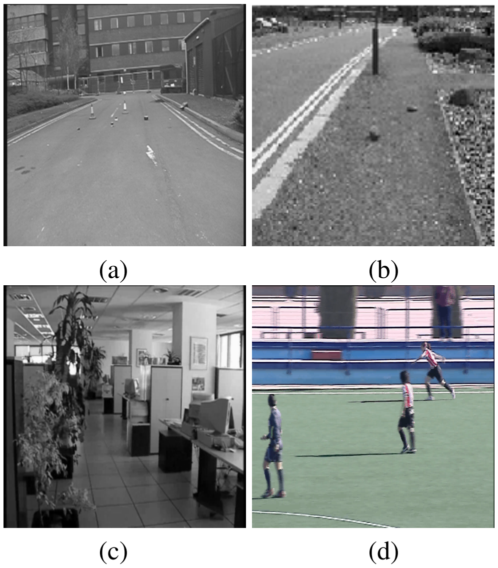
generalised as $f(\theta, \mathbf{r}) = \sum_{i=1}^{N} \omega_i \mathcal{N}(\mathbf{x}_i, \theta)$, where ω_i are weight factors and will be determined when we carry out similarity measurements. Based on the Jensen's inequality, we have an alternative object function as $\log f(\theta, \mathbf{r}) \geq \sum_{i=1}^{N} \log \left(\frac{\omega_i \mathcal{N}(\mathbf{x}_i, \theta, \mathbf{r})}{q_i} \right)^{q_i}$, where q_i is a non-negative constant that satisfies $\sum_{i=1}^{N} q_i = 1$.

Considering the current estimation θ_k and $\mathbf{r}_k$ (k indicates current state), we iterate the following E and M stages via the expectation-maximisation (EM) algorithm [4]:

(1) E-stage: Assuming that θ_k and $\mathbf{r}_k$ are fixed, we expect to obtain q_i that maximises the right hand side of the object function. The solution is expressed as: $q_i = \frac{\omega_i \mathcal{N}(\mathbf{x}_i, \theta_k, \mathbf{r}_k)}{\sum_{i=1}^{N} \omega_i \mathcal{N}(\mathbf{x}_i, \theta_k, \mathbf{r}_k)}$.

(2) M-stage: Considering q_i as constants, we maximise the right side of the object function with respect to θ and $\mathbf{r}$. The inlier fraction θ is solved by $\theta_{k+1} = \frac{\sum_{i=1}^{N} \mathbf{x}_i \omega_i \mathcal{N}(\mathbf{x}_i, \theta_k, \mathbf{r}_k)}{\sum_{i=1}^{N} \omega_i \mathcal{N}(\mathbf{x}_i, \theta_k, \mathbf{r}_k)}$, where $\mathbf{r}$ is updated according to the following equation $\mathbf{r}_{k+1} \propto \sum_{i=1}^{N} q_i (\mathbf{x}_i - \theta_k)(\mathbf{x}_i - \theta_k)^T$. This E-M iteration will terminate if and only if $|\bar{\theta}_{m+1} - \bar{\theta}_m|$ is less than a pre-defined threshold ($\bar{\theta}_m$ denotes an averaged θ in group m). In other words, the difference between two distributions instead of two consecutive samples is used as a stopping criterion.

Fig. 3 illustrates the estimated ground planes, highlighted by red color, using two different techniques. It is observed that the proposed scheme leads to more accurate coplanar determination. For example, Fig. 3(a) shows that the points on the stones (in the image centre) have been correctly identified to be over the ground plane by the proposed approach. At the same time, the classical RANSAC plane fitting approach fails to do so (Fig. 3(b)).



(a) (b)

(c) (d)

Figure 4. Four test sequences used in this paper.

4. Experimental work

We conduct experiments to demonstrate how effectively the iterative RANSAC plane fitting scheme works in extracting flat surfaces, particularly ground planes. The performance of the proposed method is compared to that of the classical RANSAC plane fitting scheme with the constraint where the number of outliers falls below a predefined threshold. Four image sequences have been tested and their example frames are illustrated in Fig. 4.

Fig. 5 illustrates two neighboring image frames of a test sequence namely "campus", superimposed by the detected corner features (see Fig. 5(a) and (b)). It exhibits in Fig. 5(c) and (d) that the proposed RANSAC plane fitting scheme results in better outcomes of flat surface fitting. For example, Fig. 5(c) shows that using the proposed method we are able to correctly identify most points on the ground. Fig. 5(d) denotes a significant number of points on the buildings have been incorrectly classified to be on the ground plane by the classical technique. Meanwhile, the points on the ground plane shown on Fig. 5(d) are less dense than those of Fig. 5(c), which remains an issue in the classical method.

As an exmaple, Fig. 6 presents statistical outcomes of the averaged Euclidean distance between the overall 3D points (from depth maps) and the fitted flat surface by the proposed and classical plane fitting schemes. Smaller Euclidean distance indicates better accuracy. For justification purpose, we here reveal partial statistical results of the im-

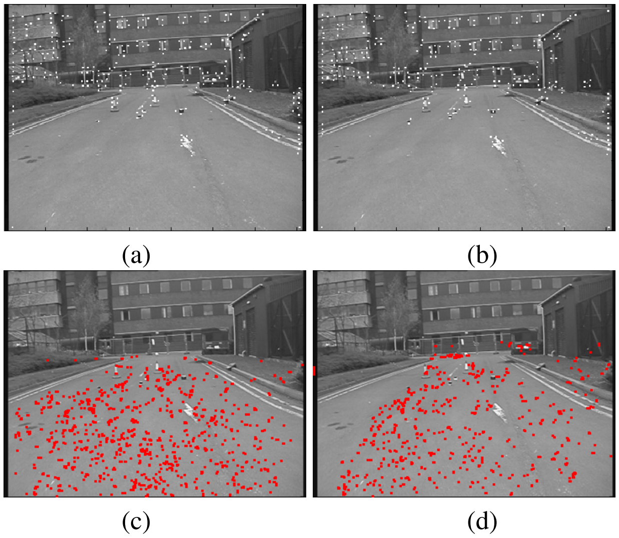


Figure 5. Examples of the estimated ground plane in sequence "campus" by two different methods: (a) and (b) feature extraction, (c) outcome of the proposed method, and (d) outcome of the classical method.

age sequence shown in Fig. 5, where ten consecutive image frames are employed for analysis in a pair-wise style. We observe that the proposed plane fitting scheme holds better accuracy of plane fitting than the classical one due to maximum likelihood estimation.

5. Conclusions and future work

We have described a technique for effective recovery of flat surfaces from digital video content. The developed system has been evaluated in a vast number of experiments. Flat surfaces are extracted from the scenes by analysing the video content, e.g. correspondence and 3D recovery. A novel iterative RANSAC plane fitting scheme was proposed. We conducted experiments of retrieving flat surfaces from videos, and the results confirmed that our plane retrieval technique was more accurate than the classical method. To develop an algorithm working for uncalibrated scenes, on-line camera calibration will be integrated into the proposed plane fitting platform. Also, more senarios need to be evaluated in the future.

Acknowledgement

This work was supported by European Commission under Grant FP6-045189-STREP (RUSHES).

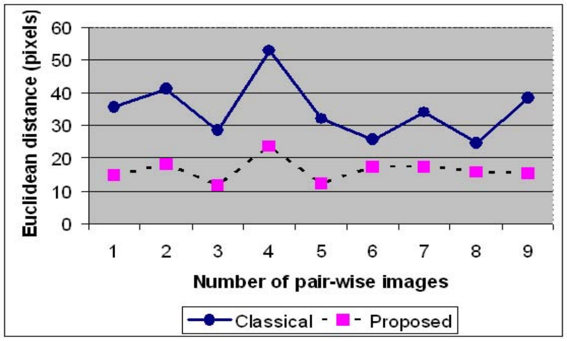


Figure 6. Illustration of the averaged Euclidean distance between the recovered 3D points and the fitted planes in sequence "campus".

References

[1] Rushes project deliverable d5, requirement analysis and use-cases definition for professional content creators or providers and home-users. In *http://www.rushes-project.eu/upload/Deliverables/D5_WP1_ETB_v04.pdf*, August 2007.

[2] A. Bartoli and P. Sturm. Structure-from-motion using lines: Representation, triangulation, and bundle adjustment. *Computer Vision and Image Understanding*, 100(3):416–441, December 2005.

[3] M. Davis. An iconic visual language for video annotation. In *Proc. of IEEE Symposium on Visual Language*, pages 196–202, 1993.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, Series B 1977.

[5] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24:381–395, 1988.

[6] C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. of Alvey Vision Conference*, pages 47–152, 1988.

[7] R. Ohbuchi and J. Kobayashi. Unsupervised learning from a corpus for shape-based 3d model retrieval. In *Proc. of the 8th ACM international workshop on Multimedia information retrieval*, pages 163–172, New York, NY, USA, 2006.

[8] J. Sivic and A. Zisserman. Video google: a text retrieval approach to obejct matching in videos. In *Proc. of Ninth IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[9] J. Tangelder and R. Veltkamp. A survey of content based 3d shape retrieval methods. In *Proc. of International Conference on Shape Modeling*, pages 145–156, 2004.

[10] H. Zhou, A. Wallace, and P. Green. A multistage filtering technique to detect hazards on the ground plane. *Pattern Recognition Letters*, 24(9-10):1453–1461, 2003.